Aalto University
School of Business

MACHINE LEARNING MODELING FOR REAL ESTATE

Predicting residential property prices in Oulu, Finland

Bachelor's Thesis
Ville Karlström
Aalto University School of Business
Information and Service Management
Fall 2020

Author Ville Karlström

Title of thesis Machine learning modeling for real estate

Degree Bachelor's degree

Degree programme Information and Service Management

Thesis advisor(s) Andrei Vedernikov

Year of approval 2020 **Number of pages** 25+5 **Language** English

Abstract

The real estate industry has long been relying on simple valuation methods and heuristics. The most commonly used methods in the industry are Comparable transactions and NOI-method. The rise of machine learning and big data are, however on the verge of changing that. Large companies are more and more investing in intelligent and automated valuation methods owing to their speed and cheapness compared to more traditional methods.

In this study, the Finnish housing market in the city of Oulu is considered for implementing machine learning algorithms to property data to predict prices.
The goal of the study was to find evidence on the usefulness of machine learning models for property price prediction. For identifying meaningful results several different algorithms were implemented.

The results show that even on a small amount of data, machine learning algorithms can produce promising results and more accurate ones than those produced by traditional methods. The average error was 10.87% on the machine learning model compared to 12 % typically produced by real estate appraisers.

The most accurate and robust algorithm was the decision-tree-based ensemble model Random forest.

Keywords Machine learning, Big data, Real estate, Valuation

Table of Contents

1 Introduction

1.1 Motivation

The utilization of big data and machine learning is projected to grow rapidly in the coming years. An article by Forbes (2020) estimates that the global machine learning market will increase by as much as 44% per year over the next five years. The market value of machine learning is projected to reach 20 billion US dollars by the end of 2024. This is going to revolutionize several business areas and reshape the way people work. As a result of big data and machine learning, old industries will disappear, and new ones will emerge.

According to a report (2018) by management consultancy firm McKinsey Co., one of the biggest industries facing a major change due to big data is the real estate sector.

Real estate is the most valuable asset class in the world measured by market value (Kok et al., 2017). For example, in the US real estate accounts around half of the American households' overall net wealth (Wolfgang Breuer, 2020). The impact of real estate on the economy is therefore considered to be significant.

However, compared to other investment categories, the utilization of algorithms and data is still very much in its infancy. Typically, in the real estate industry, the investment process is often based on hedonistic pricing models combining only the utmost basic features on every property (Kok et al., (2017) & Pagourti et al. (2003)).

Due to the explosive growth of machine learning and big data in the coming years, understanding them will become very important in the future in every segment of real estate. The benefits of machine learning methods will especially affect real estate valuation methods. Automated methods are cheaper and much faster than traditional methods, which will greatly enhance the valuation processes (Kok et al., 2017). According to the McKinsey report (2018), property returns can be estimated with an accuracy of 50 % on average, with machine learning algorithms this accuracy can be as high as 95%. With that large accuracy improvements, real estate companies cannot ignore utilizing these methods. There will be a significant increase in demand for skills and knowledge in machine learning and big data for real estate industry.

1.2 Research design

In this thesis, I will look at the potential benefits of machine learning and big data in property price modeling, especially how it can improve value appraisals and price prediction.

The title and subtitle of my thesis are:

Machine learning modeling for real estate
Predicting property prices in Oulu, Finland

To study the subject, I have formed three research questions to be answered to gain knowledge on the subject.

1. What are the current methods used for a property valuation?

2. What algorithms are the most suitable for property price predicting?

- Measured by certain Key Performance Indicators

3. Can machine learning aid in property valuation?

- What features are the best to predict prices?

- Are algorithms more accurate than traditional appraisal methods?

With the help of research questions, I will be examining the benefits of machine learning algorithms in real estate appraisal. To gain more understanding of the topic, I am developing a machine learning model that utilizes real data on the housing market from a Finnish city Oulu.

The programming done in this study is conducted using Python programming language. Data collecting is done with the Python package BeautifulSoup, data cleaning, and feature engineering with Pandas and NumPy. The modeling and evaluation parts were done with Scikit-learn. For the data storing SQL database was created and used.

1.3 The structure of the thesis

The structure of my thesis is presented as follows: Chapter 1 will introduce the thesis topic, structure, and research questions. Chapter 2 will present a literature review of previous research to find information on current methods on real estate valuation and the different machine learning solutions already implemented on real estate. The literature review in chapter 2 will answer research question 1 and partly to question 2. Chapters 3 and 4 describe the process of collecting the data and implementing a machine learning model on the data. Chapters 3 and 4 will find answers to research question 2. Chapter 5 examines the models' performance and the benefits of machine learning modeling, using the results obtained in Chapters 3 and 4. Chapter 5 will answer research question 3. Chapter 6 provides ways to improve the model and discusses the limitations of this study.

I have narrowed down my thesis to include only residential properties. The advantage of residential properties from a research point of view is better availability and amount of data, compared to e.g., office properties. Before modeling, the data used in the thesis needs to be cleaned up and modified for more accurate modeling results. The data consist of records of real estate for sale in the Finnish city Oulu.

For this study, I have chosen to use data on properties for sale instead of actual sales data, which would be available from the Real Estate Agency Federation. There are two reasons to use data on properties for sale. First is the number of features stored for each record. Actual sale data stores fewer features for every record. By accessing for sale data, it is possible to gain access to a lot more features for every record. The second reason is time. In the existing time constraints for the study, it was faster to code a program to collect the data than it would have been to request it from the Real Estate Agency Federation. Although for a more comprehensive study with a broader timeframe (e.g., a Master's thesis) collecting actual sales data from a longer time series would be reasonable.

2 Literature review

2.1 Real estate valuation

The task of real estate appraisal is to provide an estimate of the potential market price of the property. The accuracy of the valuation can be assessed by comparing the given estimate with the actual purchase price. A real estate appraisal is asked by several actors. Typically, banks, lenders, and investors are most interested in appraisals. Banks and lenders make use of assessments when making decisions about the profitability of the target property and its collateral valuation. The valuation of collateral is one of the most important factors in determining the risk of the investment. The more accurately a bank or a lender can assess the real value of collateral, the lower the risk of borrowing money. Lower risk is reflected in a lower risk premium for the investment measured by interest rate. Investors need assessments to make better investment decisions. The better an investor can estimate the true value of his investment, the more accurately an investor can predict his future cash flow. This leads to lower risk, which eventually results in lower risk-premiums for the investment.

Real estate appraisals are usually conducted either by in-house managers or third-party appraisals or consultants in exchange for a fee. Typically, an appraisal charges 3,000–5,000 $ and takes around three to four weeks with an average error ranging from 10% to 15% (Kok et al., 2017). Cannon and Cole (2011) studied the appraisals from 1984 to 2010. They estimated that the error of a typical property appraiser is about 12%, either below or above the realized price. Cannon and Cole (2011) also noticed appraisers tend to make even worse estimates on-peak market cycles, overestimating real estate values in the downturn and underestimating in the upturns. Their results were consistent with Fisher et al. (1999), who found that on average real estate appraisals were 9% -12,5% wrong on a 20-year-period. Cannon and Cole (2011) also discovered differences between in-house-appraisals and third-party-appraisals. In-house appraisals were performing worse than external ones. Although in-house appraisals were still more accurate than the ones without an appraisal.

2.2 Traditional appraisal methods

There are numerous different methods and models for real estate appraisal. Typically, however, evaluations are performed using a few of the most common methods. Next, I will introduce the most common valuation methods.

2.2.1 Comparable transactions

The most common method for valuation is to compare the price with the prices of comparable trades made in the area in the past. In this case, the price of the property is thought to correspond

to the price previously paid for a similar property. This method is known as the comparable method. When using the comparable method, the appraisers must make adjustments to subject property since in very limited cases are two different properties identical (Pagourti et al., 2003). Appraisers look for differences in property characteristics e.g., room-count, floor-number, or difference in view. The challenge for the comparable method is its heavy dependency on the available data (Pagourti et al., 2003). If the sales data is limited or not available at all, the accuracy of the comparable method can be very limited.

McCluskey et. al (1997) determined the comparability of properties by first determining the accepted distance in similarities and then calculating the distance by adding the squared differences in characteristics in the properties. The properties with the lowest values are then selected for comparison.

2.2.2 NOI-method

Another very commonly used method is the net-income-based method. In this case, the annual net return is calculated for the property, and the required rate of return is set for the investment. The division of these then gives the price of the property. The method tells the parties the price that the buyer can pay at most for the apartment and still get the required annual return on their investment (Kok et al., 2017).

The net-income-based method or NOI-method is usually most used by investors. NOI-method's advantages are in its ease of use. To implement it, one must only determine the market rent for the property and the required cap rate. If the property is not occupied at the time, investors need to estimate the potential available net operating income from the property. Usually, investors use both the NOI-method and the comparable method to determine the value of the property, since in many cases cap rates are not stable across multiple areas (Pagourti et al., 2003). This is due to rates being lower in areas where the demand for properties is higher. So, for investors, a good heuristic is usually to filter areas where the cap rate meets their required return and then determine the value of the property using the comparable method. This way investors are not so likely to end up overpaying for properties.

Even though the NOI-method used together with the comparable method can provide a solid base for valuation of the underlying asset they are still highly prone to errors. NOI-model relies on two assumptions: Firstly, the net operating income can be measured accurately for perpetuity. Secondly, an appropriate rate of return can be calculated. Still, research shows that appraisers systematically overestimate estimated NOI (Öhman et al. (2011)). This sets the NOI-method in questionable light since it can lead to systematic overestimation in property values.

2.2.3 Special situations

Other valuation methods include e.g., the Development method and the Contractors method (Pagourti et al., 2003). These methods are typically used for special situations which include for example buying an unfinished building or building that is so special that it has no meaningful comparisons available. In these situations, the valuation of property becomes more customized since factors influencing the price increase considerably. This also reflects the uncertainty of the appraisal increasing the variance between the predicted value and the real market value.

2.3 Big data and machine learning methods

Machine learning means creating models that take advantage of artificial intelligence to improve their performance by experience without explicitly told so from outside (Mitchell, 1997). Usually, machine learning is divided into three different paradigms which are supervised learning, unsupervised learning, and reinforcement learning. Paradigms represent different solutions to problems in which machine learning can be utilized (Mitchell, (1997) & Geoffrey E. Hinton, (1999)).

Machine learning models typically go through a creating process consisting of data preprocessing, learning, and evaluation phases. The purpose of preprocessing is to modify the data to be more suitable for machine learning algorithms since data is usually unstructured, messy, and inconsistent (Lina Zhoua, 2017). In the learning phase model selection and parameter tuning are conducted to produce the desired predictions (Lina Zhoua, 2017). Lastly, model predictions are evaluated with different evaluation frameworks. Frameworks can consist of multiple accuracy metrics, statistical tests, or error estimation measures (Lina Zhoua, 2017).

2.3.1 Algorithms

Multiple different algorithms can be utilized in solving problems with machine learning. In this study, I will be taking a look into two widely used algorithms in regression tasks: Linear regression and Decision trees.

Linear regression

Linear regression is divided into three categories: simple linear regression, multiple linear regression, and multivariate linear regression depending on the input and output variables. Property price predicting is considered multiple linear regression since there are many input variables and one output variable. The idea of linear regression is to predict the output variable (y) based on certain input variables (x) by formulating a function with individual weights for each input variable. The linear regression model is formed according to the equation below:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots b_nX_n$$

$$Where\ Y = the\ dependent\ output\ variable$$

$$b_0 = constant\ (intercept)$$

$$b_1 \dots b_n = coefficents\ of\ X_1 \dots X_n$$

$$X_1 \dots X_n = independent\ input\ variables$$

Equation 1. Linear regression

Linear regression is good for finding different relationships between multiple independent and dependent variables. Linear regression is established by fitting a line to data that minimizes the squared residuals produced around the line. From the R2-score formula it can be seen that correlation ratio and fitted linear model have an inverse relationship: lower the residuals, higher the R2-score. If the line fits perfectly on observed data values, then the squared residuals are equal to actual squared values of the data producing an R2-score of 1.

According to Kok et al. (2017), the benefits of using linear regression algorithms are in their ease of use and fast implementation. The limitations of linear models arise when the models become more complicated and the variables contain less linearity. Like Kok et al., (2017) point out, the main problem in the linear model is the enlargement of the consistent formula to the whole data spectrum.

Decision trees

Decision tree algorithms are algorithms following the principle of a decision tree. The decision tree is a way to illustrate different scenarios and the different potential outcomes it can have. Trees are built up from nodes and splits. A node represents instances or 'tests' and splits the different outcomes it may have, e.g., coin flip (node) can have two outcomes heads or tails (splits). Decision trees can make decisions about nodes and splits based on different rules. Breiman (1984) introduced the reduction in variance method for regression tree building. Reduction in variance is based on the rule, where each split is decided based on minimizing the variance, the node has on that split. Variance is calculated as follows (2020):

$$Variance = \frac{\sum(X - \mu)^2}{N}$$

Equation 2. Variance

The decision tree has multiple advantages over the linear model. Firstly, decision tree algorithms do not require the data to be linearly dependent. This way decision trees can produce accurate predictions even though the data might not contain direct linear relationships. Since on many occasions that is the case, decision trees are more suitable for multiple problem-solving tasks compared to the linear model. The second major advantage is the decision tree's capability to

handle categorical variables (Witten et al., 2013). Since linear models can handle only numerical and binary variables, the ability to also analyze the categorical variables increases the probability of better model performance. Thirdly, decision trees handle collinearity better (Nam, 2019). Collinearity results from different independent variables being highly correlated with each other, which negatively affects the model's performance.

The drawback of the decision tree model is its propensity to overfit the data. Since decision trees tend to find relationships from the data without linearity, it usually ends up overfitting the model (Bramer, 2007).

2.4 Applications for real estate valuation

Most academic research papers have framed asset valuation problems under supervised machine learning problems (Ndikum, 2020). Supervised learning means giving the model different output-input pairs to train and to find patterns from them, from which the model will then create a function to predict the results from new data without the output label.

Traditional real estate appraisal methods are often limited in terms of the variables used. Normally, an appraiser is only able to consider a maximum of 10–15 different variables on a property on which an appraiser must base one's price estimate (Kok et al., 2017). This limit is not restricting machine learning models. Theoretically, those models can consider an infinite number of variables. Machine learning models can also be taught to recognize patterns in almost everything.

Since real estate properties contain multiple linear variables, as well as non-linear variables the algorithm must be able to handle both of these types. Real estate data will also probably contain overlapping features with multicollinearity (e.g., rooms and square meters). Based on these two constraints tree-based algorithms are more suitable for this study. Linear regression will however act as a benchmark to compare the models' performances. For the tree-model to prevent overfitting and to enhance performance using ensembled methods is appropriate.

3 Methodology

3.1 Theoretical framework

Machine learning model implementation into the real estate appraisal process consists of multiple steps. Singh (2020) introduced a framework for model implementation, which consists of five steps (figure 1.):

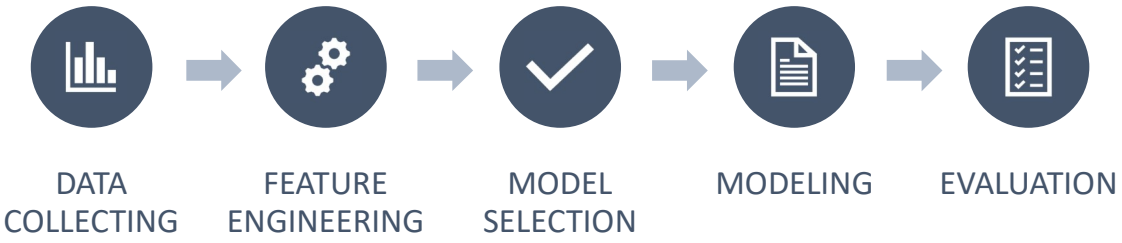


Figure 1. Theoretical framework of machine learning model implementation

The purpose of this study is to find evidence of the benefits of machine learning algorithms on property pricing. To establish a reusable process, it is necessary to define a suitable theoretical framework to illustrate the key elements of the process. The framework introduced by Singh (2020) fits these needs in its simplicity, still covering all the important elements.

3.2 Data collecting

As a data collection method, I web scraped etuovi.com, which is the most popular portal for selling properties in Finland. Web scraping means collecting the data openly available on the website and storing it for further inspection. Every website is built on HTML-code. HTML-code can be retrieved by making a GET-requests to the server. By examining the code, it is possible to see different elements from the site. Elements represent values and tags shown on the website. Inside elements in the HTML-code specific values can be found e.g., price of the property or building year. I built the web scraper to first extract property IDs from the property listing with specific search parameters (city name). After that, I iterated through those IDs to find certain features for each property. When web scraping a website, it is important not to disturb the site too much which means sending too many GET-requests in a short period as this can cause too much traffic to the server and eventually slow down the website. Whenever the scraper accesses property features via

ID, it sends a GET-request to the server. To be polite to the website, I developed a timer between requests, basically telling the scraper to keep a small break (2-3 seconds) every five properties.

The data set was collected on November 10th, 2020 on the apartments in the city of Oulu from etuovi.com. At the time, there were 2,164 apartments for sale in Oulu. After the scrape, the data had to be cleaned up, as several records were missing relevant data. I also decided to exclude properties that were not part of housing cooperatives (detached houses) because the variability of the different characteristics of these properties were unnecessarily large for this study. The final number of records in the data set was 1,813.

3.2.1 Features

Web scraper gathered 12 features for every property (Table 1). Features included information on size, location, property characteristics, and price. Some important features were also left out. The scraper did not get the information on balconies, housing associations loans, or nearby services. These were due to these features being inconsistently presented in the HTML-code in other words the scraper was not able to retrieve that information.

Features can be of a certain type e.g., categorical or numerical. Categorical values mean they can have two or more different values, but the values cannot be placed in an order in any way. Numerical values refer to integer or decimal values.

Table 1. Collected features for every record in the data and their types

Feature	***Type***
Property ID	Numerical
Address	String
House type	Categorical
Rooms	Numerical
Square meters	Numerical
Floor	Numerical
Maintenance fee	Numerical
Sauna	Categorical
Plot ownership	Categorical
Building year	Numerical
Price	Numerical

3.3 Feature engineering and data cleaning

Feature engineering refers to modifying variables derived from data to be more suitable for a machine learning algorithm. Feature engineering is an important step in modeling, as it has a significant impact on the quality of the predictions produced by the model (Zheng, 2018).

Before modifying the data to a more appropriate format, data can be visualized to identify its key characteristics and to get a better understanding of the data type (figure 2.). Visualizing the data also shows if any features contain abnormal values that are for example too high or too low for them to be reasonable.

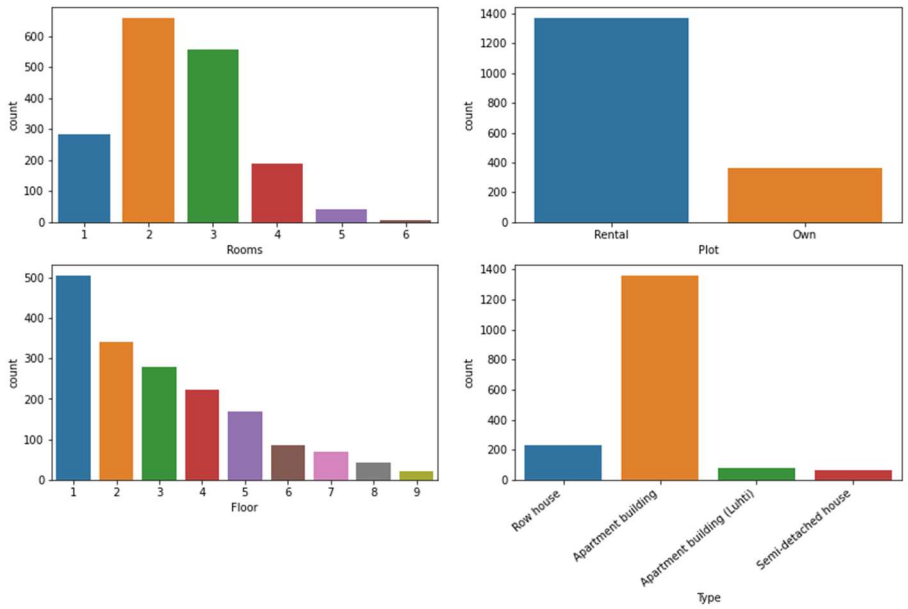


Figure 2. Visualization of data records

At first, it can be seen that the majority of records are from apartment buildings, and most buildings are located on rental plots. The most common room counts are 2 and 3, and the most common floor is the first floor. According to statistics of Finland distribution between Finnish apartment building room types are as follows 9% (16%) 1 room, 41% (36%) 2 rooms, 34% (30%) 3 rooms and 14% (18%) 4 or more rooms. Data distribution of this study was marked in parentheses. From that distribution, it can be seen that the data presents a somewhat good representation of Finnish housing distribution.

For the data to be ready to be assessed to different algorithms it needs to be cleaned and formatted correctly. First, I cleaned the data set by removing properties that did not have price data or the maintenance charge data, since the amount of them was relatively low and they were outliers in the data. Outliers are significant differentiation from the other data. They could be due to variability in the measurement or an error in the model (Grubbs, 1969).

Price data can be visualized (figure 3.) with a distribution plot to see if some records should be examined more thoroughly.

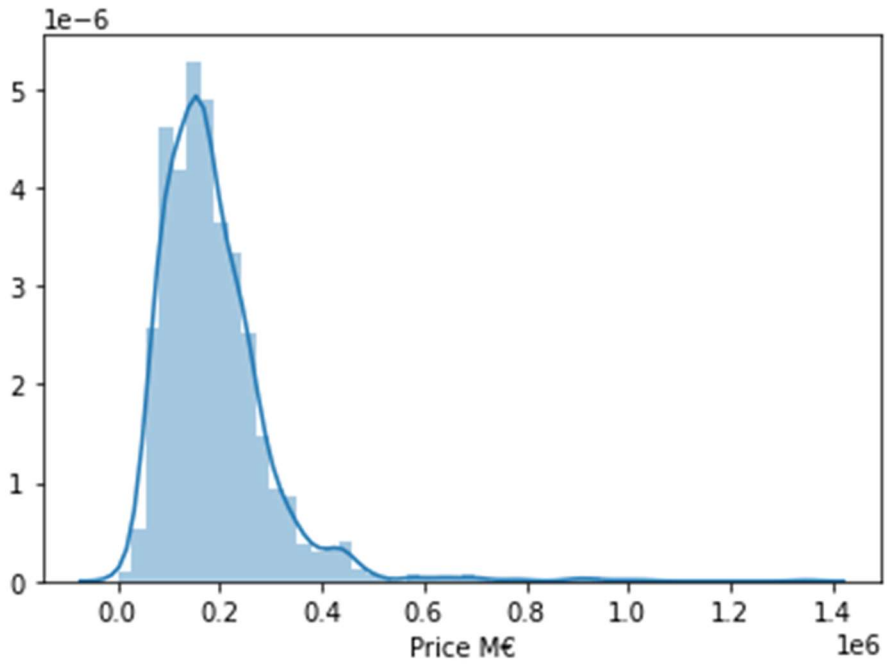


Figure 3. Distribution of Property Prices

The distribution plot shows the data is a little skewed with a long right tail. It is still hard to see if those records on the right tail are outliers, and where to draw the line in the price when determining outliers. To further examine the outliers the data can be visualized with a boxplot (figure 4.) to see if higher-priced apartments should be considered as outliers. Boxplot divides the observations into quartiles similarly to normal distribution. The boxplot categorizes observations differing over 1.5 times interquartile as outliers (2020).

$$IQC = Q3 - Q1$$

$$Outlier = 1.5 * IQC$$

Equation 3. The formula for calculating outlier values

The boxplot shows the data containing some outliers on the higher end of the price spectrum. To deal with outliers all properties with a price of over 500,000€ will be removed. With these data cleaning steps, the dataset ends up with 1782 records.

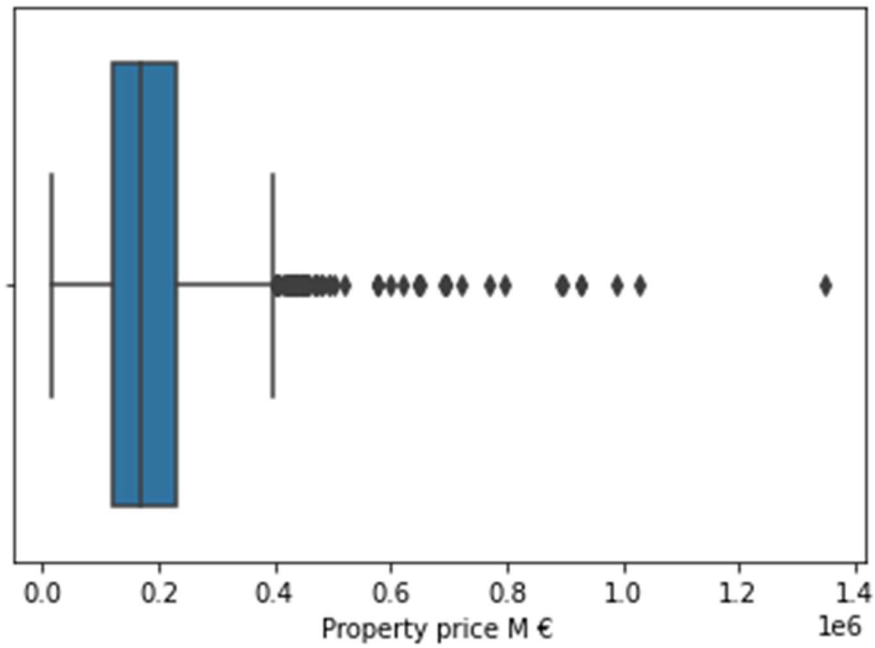


Figure 4. Distribution of Property Prices

After examining the data and removing outliers the next step is to form a feature correlation matrix. From the correlation matrix, it can be seen if the data consist of any direct linear relationships between different features (figure 5.). The correlation matrix forms correlation coefficients between different numerical features in the data. Correlation is computed using the Pearson correlation which cannot handle categorical features unless they are in binary 0/1 form.

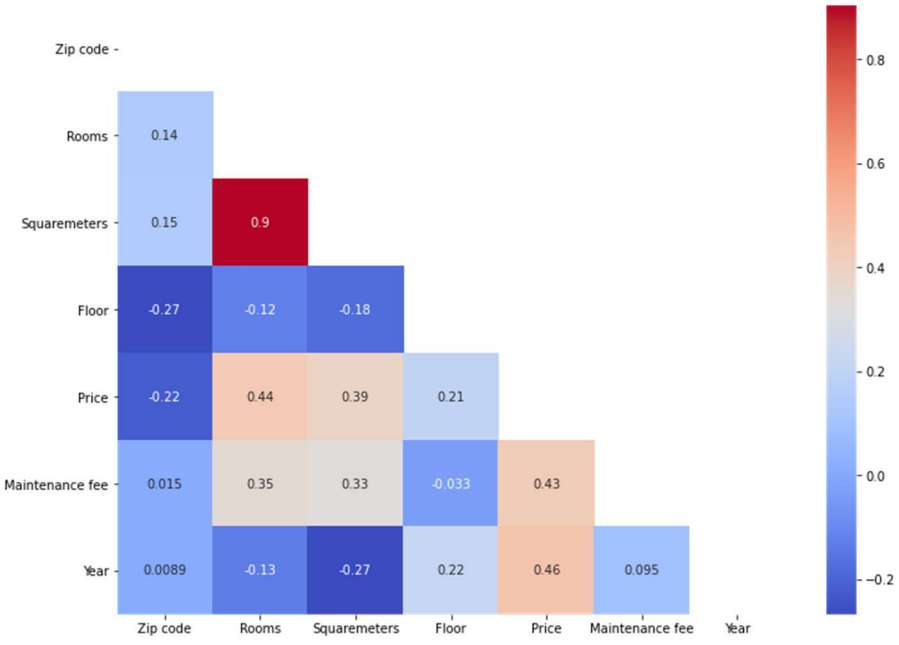


Figure 5. Correlation between numerical features

The correlation matrix shows that variable price has the most correlation with square meters, maintenance fee, room count, and building year. All of them have correlations with prices between 0.39 - 0.46. This could indicate their role as being important in predicting the price. Despite visible relationships, correlations are not clear indicators of feature importance. For example, the correlation between price and zip code is almost neutral (-0.062). Still, it is fairly certain that on average apartments are more expensive in lower zip code areas e.g., on the central vs. in the suburbs.

To increase the quality of features, the distributions of feature variables need to be examined. The distribution of values is an important factor since some machine learning models tend to perform better with normally distributed data (Frees, 2007). Normalization is not necessary for this study due to model selection, but it helps models process the data faster, making the modeling part faster.

To remove skewness from data and to make the data closer to the normal distribution, some scaling operations for the data can be applied. For numerical features (excluding building year and zip code) in the data, a Box-Cox transformation is used to remove skewness and to make them normally distributed. Box-Cox formula is defined as follows:

$$y^{\lambda} = \begin{cases} \frac{y^{\lambda} - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log y, & if\ \lambda = 0 \end{cases}$$

Equation 4. Box-Cox transform formula

Feature distribution can be seen to be closer to normal distribution after box-cox transformation by examining the distribution of maintenance fee before the transformation (left) and afterward (right) (figure 6.).

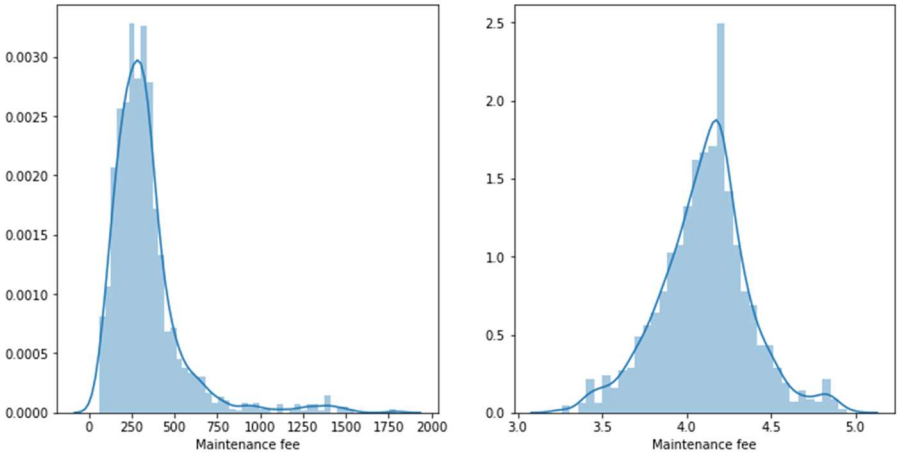


Figure 6. Effect of Box-Cox transform. Distributions of maintenance fees before (left) and after (right)

To make the building year more suitable for the model, I will divide building years into ten bins. Each bin will be approximately 10 years. So, e.g., an apartment built in 1950 would be placed in a bin consisting years e.g., 1945 – 1955. Binning can aid in model tuning by categorizing discrete numerical variables.

Lastly, before the data set is ready for modeling, categorical features are turned into dummy variables. A dummy variable means creating an additional column as a binary representation from every different value from the feature (Sharma, 2003). E.g., If the property can have a sauna or not, the data would be altered to contain two columns holding value 1 if true and 0 if false. Final dataset contains 1782 records and 65 features of which 53 are dummy features.

3.4 Model training and testing

Supervised machine learning models need to be trained before the prediction phase. Training means providing the model data with all the feature values and label values for those features. The model will look for patterns and linear relationships from the training data. After the training, the model is given fresh unseen data with all the same features but without the label value. Then from the test data model will try to estimate a label value for every record in the data. In my modeling, I will be splitting the dataset into train data and test data with a random 75/25 split. The train data contains 1,336 records and the test data 446 records.

When training the model, it is important to make sure the model does not become overfitted. Overfitting refers to a situation in which the model can be extremely accurate when looked at its performance on finding relationships from the training data, but when seen new fresh data the model performs poorly (Claesen & De Moor, 2015). Overfitting is typical when trying to get the model as robust as possible.

The opposite of overfitting is underfitting. Underfitting means the model performing poorly both on the train and test data. This balancing is often referred to as bias-variance trade-off (Claesen & De Moor, 2015). To prevent overfitting certain steps can be taken. Firstly, I have kept the number of variables relatively low, while keeping an eye on the quality of variables. Secondly, I implemented a cross-validation algorithm to model training and testing phase. Cross-validation creates random samples of the data and tests different models on them trying to see if models are overfitted. Thirdly, two of the three models (Gradient and Random forest) are ensembled models. Ensembled models are built by combining multiple learning algorithms to produce better results. Random forest e.g., is created by bagging multiple decision tree algorithms together and run parallel to get more accurate results and to reduce variance in models.

Training performance can be enhanced by tuning the parameters for each model. These hyperparameters determine the lengthiness and deepness of the training process. Hyperparameter refers to models' parameters that have been determined before training the model e.g., number of nodes in the decision tree. Hyperparameter tuning is usually done manually by running the model multiple times and changing the parameters between rounds to find the optimal combination (Claesen & De Moor, 2015). In this study, I will try to find the optimal parameters by minimizing the mean absolute error (MAE) while maximizing the R2-score in predictions.

3.5 Key Performance Indicators

To assess the quality of the machine learning model certain performance measuring metrics, need to be added. There are several different metrics for measuring the quality of the models. As performance measurement is not straightforward, it is preferable to choose more than one metric for evaluation of the models. Some metrics like the R2-score measure the model's robustness and other metrics e.g., mean squared error measure the accuracy the model produces (Kok et al., 2017). Robustness refers to the model's capability to explain the variance between predictions and actual values.

I have chosen four performance indicators for model evaluation. R2-score to evaluate model's robustness, Root mean squared error (RMSE) and mean absolute error (MAE) to predict the accuracy of the model and mean absolute percentage error (MAPE) to clarify the scale and to compare results to previous studies. The reason to pick two different accuracy metrics (MAE, RMSE) is their different handling of incorrect predictions. Root-mean squared error is a square-based measure, so it gives more weight to large prediction errors. MAE only measures the errors arithmetic mean. This means that if RMSE is meaningfully higher, predictions probably contain few very large errors but are on average modest.

R2-score is beneficial to compare the models between each other. The other metrics (MAE, RMSE, MAPE) provide ways to easily compare the actual values to predicted ones.

$$R^2 = 1 - \left(\frac{Total\ sum\ of\ squared\ residuals}{Total\ sum\ of\ squares}\right)$$

$$RMSE = \frac{1}{N}\sum_{i=1}^{n}\sqrt{(A-F)^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|A-F|$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{A-F}{A}\right|$$

$$A = actual\ house\ value$$

$$F = forecasted\ house\ value$$

Equation 5. Key Performance Indicator formulas

4 Modeling

Based on the findings made in the literature review, I chose three different models to predict house prices. Models are Linear regression, Random forest regressor, and Gradient boosting regressor. Random forest and Gradient boosting are ensembled models built by combining up multiple decision tree algorithms running in parallel to increase the performance of the model and to reduce modeling biases e.g., overfitting.

Predicting prices is a regression problem. Regression problems in machine learning refer to a situation where the aim is to predict the exact value of the label (home price) based on variables. Since property prices are continuous variables regression analysis is better suited for this study's purpose. Property prices could also be modeled by classification, by defining certain intervals for prices in which each property would be placed.

For comparison, I have also formed a fourth model: ordinary least squares (OLS) linear regression model. OLS is composed by fitting a linear model to the data without any external feature engineering to simulate the performance of a simple hedonistic model.

4.1 Ensembled models

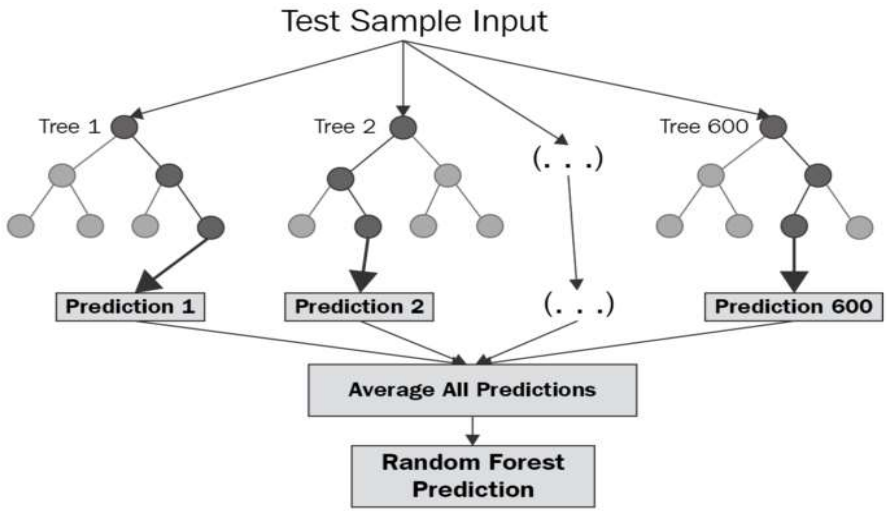
Ensembled methods are a way to improve models' accuracy and reduce variance. Ensembled method means combining multiple meta-algorithms to increase performance (Dietterich T. G., 2000). There are multiple ensembled methods of which most notable are The Bayesian methods, Bootstrap aggregating (Bagging), and Boosting. Two of the chosen models are ensembled, both of which utilize decision tree algorithms as a base algorithm and implement a meta-algorithm on top of it for better results. Random forest regressor uses the Bagging method and Gradient boosting regressor the Boosting method.

4.1.1 Random forest regressor

Random forest is an ensemble model built upon the decision tree algorithm. Random forest was created by Tim Kan Ho (1998) when he introduced the random subspace method to increase decision tree model accuracy and to reduce overfitting data features that are split into smaller subsets. Breiman (2001) introduced additions to the Random forest model by combining the random subspace method with the Bagging method developed by Breiman. The bagging (Bootstrap aggregation) method is one of the basic ensemble meta-algorithm used in machine learning to improve model results. The bagging method works by dividing training data into separate subsets.

Random forest's effectiveness is based on combining multiple low correlated decision trees. Single decision trees are highly likely to overfit the data, but since many uncorrelated trees are combined,

they produce better results with less variance. Picture 1 visualizes the principles of Random forest models.



Picture 1. Random forest algorithm principles

4.1.2 Gradient boosting regressor

Gradient boosting regressor is also an ensembled model but different from Random forest, Gradient booster utilizes boosting as its meta-algorithm. Boosting is based on an assumption of formatting strong learners from multiple weak ones (Zhou, 2012). It means combining multiple low correlating variables to formulate high correlating variables. Independent variables e.g., floor number, room count, or zip code individually have a low correlation on property price but combined have a significantly higher correlation with price.

Boosting differs from Bagging by trying to create stronger variables by changing the feature weights, whereas in bagging the model tries to produce better results by running parallel independent subversions of the original data.

Gradient boosting regressor tries to minimize the loss function produced on fitting predictions on actual data. The loss function is minimized by fitting gradient estimators to data. Loss function can be measured e.g., with a mean squared error.

$$MSE = \frac{\sum(y - \hat{y})^2}{n}$$

Equation 6. Mean squared error

4.2 Model performance

Table 2 presents the modeling results measured by chosen KPIs. From the table, it can be determined that the Random forest was the best model measured by robustness and accuracy.

Random forest had an R2-score of 0.83 with a mean average percentage error of 10.87%. Gradient booster performed slightly worse than the Random forest with an R2-score of 0.79 and MAPE 12.99%. The worst performing machine learning model was the linear regression model with an R2-score of 0.77 and MAPE of 14.65%. Still, all models outperformed the standard hedonistic model of least squares (OLS).

Root mean squared error can also be seen as being significantly higher than MAE on every model (Random forest 32817 vs. 17183). Higher RMSE compared to MAE reflects the predictions containing some larger errors, resulting in the summed square meter (RMSE) being higher. This is expected since the data contains some records that will trick the model to produce large prediction errors if it is not robust enough e.g., very large square meters with very low price. With the amount of data used in this study, it is not realistic to expect the model to cope with records like that.

Table 2. Results from property price modeling

	Key Performance Indicators			
Model	R2	MAPE	RMSE	MAE
OLS (Hedonic model)	0.66	22.39 %	48,464	32,935
Gradient Boosting	0.79	12.99 %	36,395	22,104
Random forest	**0.83**	**10.87 %**	**32,817**	**17,183**
Linear Regression	0.77	14.65 %	39,423	24,127

The results were rather unsurprising: The ensembled models overperformed over traditional linear regression, but linear regressions performance was still relatively good. When comparing the ensembled models', Random forest's performance was superior with over two percentage points accuracy edge over Gradient booster. The traditional hedonic model's performance was poor compared to other models. These findings are similar to the ones of Kok et al. (2017). They also discovered ensembled models to produce more accurate results than linear models. Also, in their study, the performance of simple OLS models was much poorer than other models. The poor performance of the OLS model indicates the importance of feature engineering and the data preprocessing process. The OLS model and the Linear regression model were otherwise similar, but the Linear regression model was supplied with feature-engineered data. This shows that appropriate feature engineering increases model performance significantly.

4.2.1 Model execution times

When evaluating the usefulness of different models, time for execution is also an important factor to consider. Different models take different times to execute, some being significantly faster than others. Execution time is directly related to the model's performance: it does not matter whether the model is accurate, if it takes too long to execute or requires too much processing power. With small amounts of data, the differences are not relevant but if models would be scaled up, the gap between execution times would get wider causing major differences in the usefulness of the models. As seen in table 3 Gradient boosting regressor was almost three times faster in training the model than Random forest. But Linear model was still the fastest which is anticipated as it is by far the simplest of the models. Results are intriguing since the Random forest was the most accurate model but by far the slowest. This raises the question of the accuracy of the models. It needs to be addressed how much processing power can be used for minor accuracy improvements.

Table 3. Model execution times

Model	***Execution time (sec)***
Linear regression	0.016
Gradient boosting regressor	0.225
Random forest regressor	0.634

4.2.2 Feature importance and Hyperparameter tuning

The figure 7. presents the most significant features for Random forest. As can be seen from the figure 7 square meters, building year, and location (zip code) were most determining for price. Those features accounted for approximately 85% of the models predicting power. Less significant features were maintenance fee, house type, floor, and room count. At first, room-counts low significance (0.86) might sound strange considering the high correlation it has with price as could be seen in figure 4. The reason is that square meters and room count are highly similar and correlated variables, so to prevent duplicated information the model ignores another variable completely.

Figure 7. shows that the building year is almost three times statistically more significant than location. This is interesting as usually the location on the property is considered the most important of all features. The reason behind the building year's significance can be in the large price gap between old and new buildings. It seems that the price gap is larger between new and old buildings than it is between centrally located and suburbs located buildings. The above could be the reason why the building year provides more information on the price to the model.

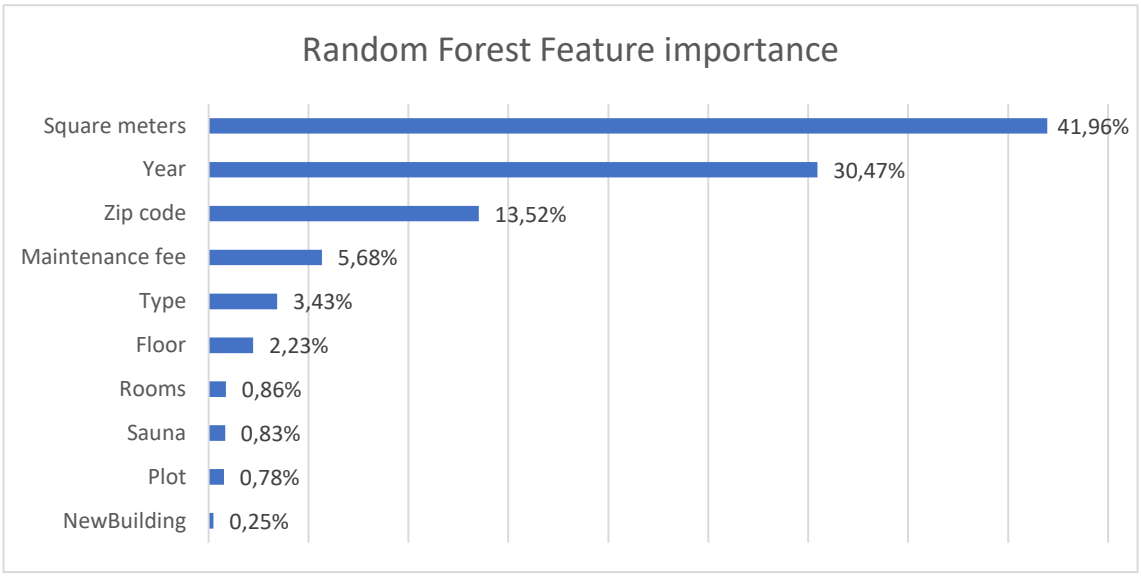


Figure 7. Feature Importance on Random forest model

To determine the optimal performance of the model several test runs are completed. To illustrate the changes each iteration has on the performance, different parameter scenarios can be examined (figure 8.). Figure 8 shows changes in price normalization and the number of estimators to make the largest effect on model performance, as the results would be 13.5 - 15.3% worse without them. Feature normalization, year binning, and location code changes can be considered feature engineering, while estimator amount and test-size scale are model parameter tuning. To find the optimal set of features and hyperparameters, multiple iterations runs have been executed.

Optimal parameters for Random forest were decision tree number of 100 and for the Gradient boosting estimator number of 100. After 100 decision trees, the model's performance did not increase anymore but the execution times grew higher since the complexity of the model grew.

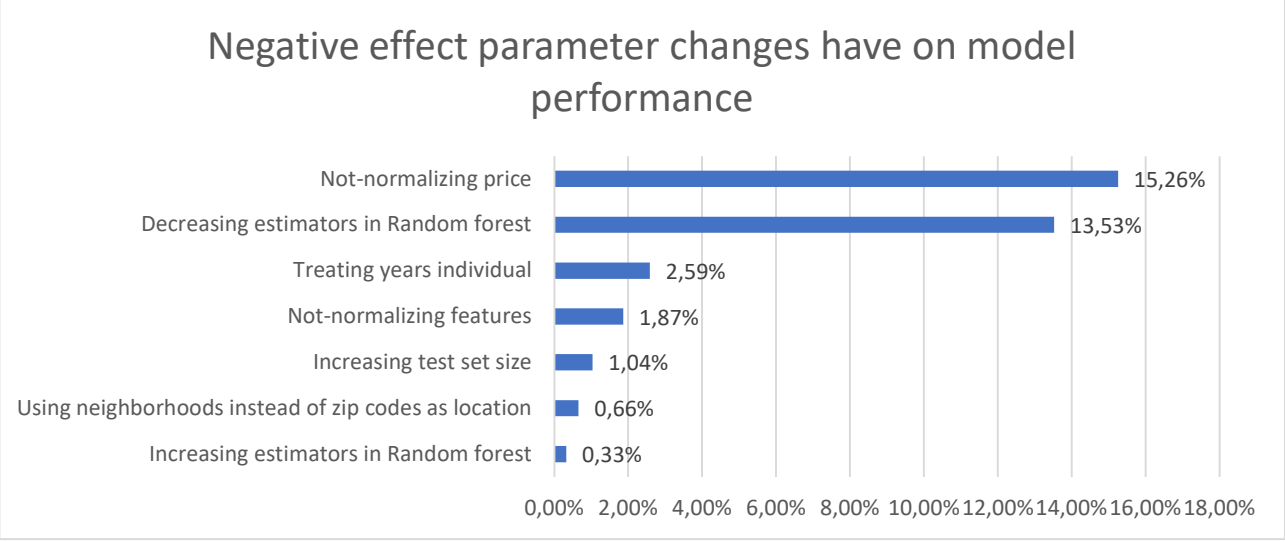


Figure 8. Parameters change effect on the Random forest model performance

5 Conclusions

To conduct this study, I have completed an analysis of previous research on machine learning and its applications in real estate valuation. From the analysis, I have chosen a theoretical framework and the appropriate models to establish the machine learning model. The analysis was conducted as a literature review. To answer research question 1 the analysis provided information on the most common models used for property valuation. Usually, properties are evaluated with simple hedonic models, the most common models being the comparable transactions and NOI-model. These models still had some drawbacks which especially emerge on weak economic cycles.

After the analysis, I web scraped property portal etuovi.com to collect property data from the Finnish city Oulu for the modeling part. The data was then cleaned and engineered for the machine learning models to process it. Data was then entered into three different models to train. After training all the models, they were supplied with fresh unseen data for predicting. Predictions were then evaluated based on previously determined Key Performance Indicators. Based on the indicators and to answer research question 2, model algorithm Random forest performed the best by all the chosen indicators.

Research by Cannon and Cole (2011) discovered the average appraisal error (MAPE) done by property appraisers to be on average 12 %. Compared to results Cannon and Cole received, results in this study sound promising. The Random forest model established in this study produced results averaging 10.87 % error in predictions compared to the actual value. Even though the model was fairly simple containing only internal features on apartments it still overperformed the manual appraisal over 1 percentage point. Based on the accuracy of the Random forest model it can be said that machine learning models aid in property valuation which answers research question 3.

When compared to previously done studies e.g., Kok et al., (2017) this study's models' performance was almost as good theirs. In their study, Kok et al. were able to create a model for price prediction with a median absolute percentage error (MdAPE) of 9.3 %. Still, the difference in performance metrics between this study and theirs needs to be addressed. Since when the skewness of the distribution is positive, *mean* value is higher than the *median* of data. In my predictions the data skewness is positive (figure 9.). This means that by using *median* instead of *mean* my results would have seemed to be slightly better.

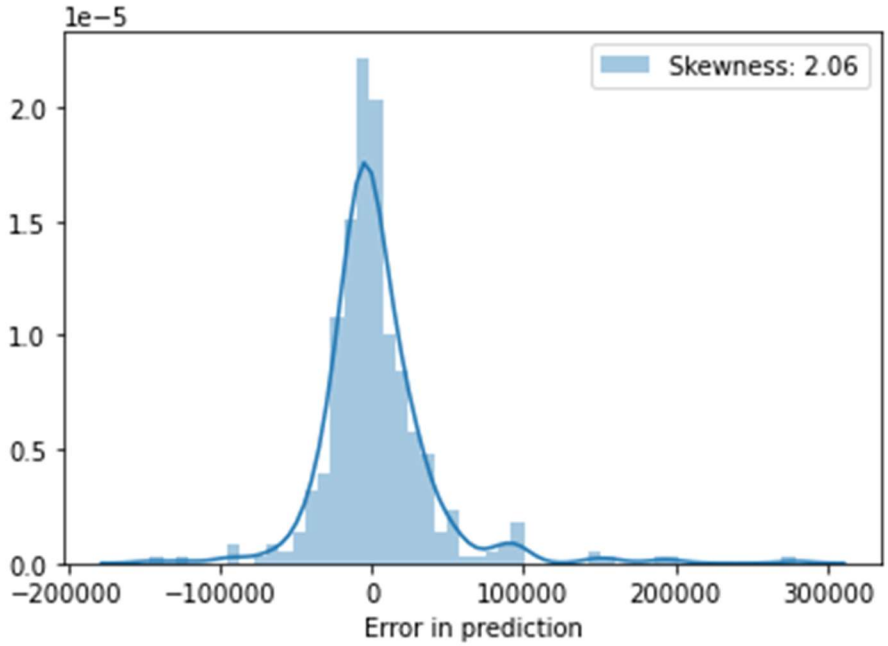


Figure 9. Prediction errors distribution

Based on these findings the potential of automated valuation tools (AVM) is significant. AVM's provide instant benefits by speeding up the valuation process with a fraction of the cost of a manual appraisal. The potential downfall of AVM's is the deviation of the predictions as some predictions contain very large errors, but on average they provide sufficient results. Since the deviation of manual appraisals is not known, models cannot be examined from that perspective.

The results were to some extent what I was expecting. After reading multiple articles and reports discussing the topic, it was rather obvious that the whole real estate industry knows the potential of AVMs and many organizations are heavily investing in the field. Still, there are not yet many concrete applications. The majority of organizations are still using machine learning only to aid in the process rather than let the algorithm make the decisions. However, the coming years will probably introduce multiple advanced applications due to the explosive growth rate of the machine learning industry.

6 Limitations and Future research

To further increase the accuracy of results for my models more data would be needed. The amount of data used in this study was very limited (less than 2,000 records), compared to Kok et al. (2017) my data was over 10 times smaller (1,752 vs 28,145 records). This strongly limits the capability of learning for the model.

The data in this study also contained only internal features of the property e.g., size, year built. To improve results, external features could be added to describe the livelihood of the neighborhoods. This would increase the model's capability to differentiate different neighborhoods, as currently the zip code is the only location differentiating feature.

For future analysis including a longer time series could potentially increase the performance of the models since it would allow other external features e.g., GDP-growth, interest rates and the number of units built/year. With time series and external neighborhood features the model would get a more in-depth view of the underlying market conditions and find more undiscovered patterns between different features and property price. To discover more complex patterns from the data, it would be a reasonable idea to implement a neural network to the model.

Still, it is good to take into consideration that the scope of this study was rather limited consisting of only one city. Since the models handled data only from one city, it is difficult to make assumptions on its performance, when assessed with multiple cities or even countries. Property markets inside cities tend to be far less heterogeneous than between cities. Adding more cities to the model would greatly increase its complexity, but also would provide valuable insights and new patterns.

7 Bibliography

4 Simple Ways to Split a Decision Tree in Machine Learning. (2020). Retrieved from www.analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/

Archana Singh, A. S. (2020). Big data analytics predicting real estate prices.

Bramer, M. (2007). *Principles of Data Mining*. Springer, London.

Breiman Leo, F. J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Breiman, L. (2001). *Random forests*.

Chakure, A. (2019, 6). *www.medium.com*. Retrieved from https://medium.com/swlh: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Claesen, M., & De Moor, B. (2015). *Hyperparameter Search in Machine Learning*. STADIUS Center for Dynamical Systems,.

Dietterich, T. (1995). Overfitting and Undercomputing in Machine Learning. *ACM Computing Surveys*.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems* (pp. 1-16). Cagliari, Italy: Springer.

Etuovi.com. (2020, 11). *Oulu asuntohaku*. Retrieved 11 10, 2020, from Etuovi: www.etuovi.com/oulu

Fisher, J. D. (1999). How Reliable Are Commercial Appraisals? Another Look. *Real Estate Finance*.

Frees, E. W. (2007). Regression and the Normal Distribution. In E. W. Frees, *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Gabriel Morgan Asaftei, Sudeep Doshi, John Means, and Aditya Sanghvi. (2018). *Getting ahead of the market: How big data is transforming real estate*. Mckinsey Co.

Geoffrey E. Hinton, . J. (1999). *Unsupervised learning: foundations of neural computation*.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*.

Ho, T. K. (1998). *The Random Subspace Method for Constructing Decision Forests*.

Holmes, A., Illowsky, B., & Dean, S. (2020). *Introductory Business Statistics*. Retrieved from https://opentextbc.ca/.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY : Springer New York : Imprint: Springer.

Lina Zhoua, S. P. (2017). *Machine learning on big data: Opportunities and challenges*.

Louis Columbus. (2020). Roundup Of Machine Learning Forecasts And Market Estimates, 2020. *Forbes*.

Mitchell, T. (1997). *Machine learning*.

Nam, H. (2019). *Predicting Diabetes Using Tree-based Methods*. Department of Statistics Uppsala University.

Ndikum, P. (2020). *Machine learning for algorithms for financial asset price forecasting*. University of Oxford.

Nils Kok, E.-L. K.-B. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *Special real estate issue*.

Pagourti, E., Assimakopoulus, V., & Hatzichristos, T. (2003). Real estate appraisal: a review of valuation methods. *Journal of property investment & finance*.

Peter Öhman, B. S. (2011). Accuracy of Swedish property appraisers' forecasts of net operating income. *Journal of Property Research* .

Sharma, S. G. (2003). *A SMART GUIDE TO DUMMY VARIABLES: FOUR APPLICATIONS AND A MACRO*.

Susanne E. Cannon, R. A. (2011). *How Accurate Are Commercial-Real-Estate Appraisals? Evidence from 25 Years of NCREIF Sales Data*. Chicago: DePaul University .

William McCluskey, W. D. (1997). Interactive application of computer assisted mass appraisal and geographic information systems. *Journal of Property Valuation & Investment*, 448-465.

Wolfgang Breuer, B. I. (2020). Recent trends in real estate research: a comparison of recent working papers and publications using machine learning algorithms. *Journal of Business Economics*.

Zheng, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientist*. O'Reilly Media.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*.